THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Optimization-based modeling of Lombard speech articulation

Supraglottal characteristics

Citation for published version:
Elie, B, Šimko, J & Turk, A 2024, 'Optimization-based modeling of Lombard speech articulation: Supraglottal characteristics', *JASA Express Letters*, vol. 4, no. 1, 015204 . https://doi.org/10.1121/10.0024364

Digital Object Identifier (DOI):
10.1121/10.0024364

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
JASA Express Letters

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 17. Nov. 2026

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Optimization-based modeling of Lombard speech articulation: Supraglottal characteristics

Benjamin Elie,[1,a] **Juraj Šimko,**[2] **and Alice Turk**[1]
[1]*Linguistics and English Language, School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, Edinburgh, Scotland, United Kingdom*
[2]*Department of Digital Humanities, Faculty of Arts, University of Helsinki, Helsinki, Finland*
benjamin.elie@ed.ac.uk, juraj.simko@helsinki.fi, a.turk@ed.ac.uk

Abstract: This paper shows that a highly simplified model of speech production based on the optimization of articulatory effort versus intelligibility can account for some observed articulatory consequences of signal-to-noise ratio. Simulations of static vowels in the presence of various background noise levels show that the model predicts articulatory and acoustic modifications of the type observed in Lombard speech. These features were obtained only when the constraint applied to articulatory effort decreases as the level of background noise increases. These results support the hypothesis that Lombard speech is listener oriented and speakers adapt their articulation in noisy environments. *© 2024 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).*

[Editor: Douglas D. O'Shaughnessy] https://doi.org/10.1121/10.0024364

Received: 25 October 2023 **Accepted:** 30 December 2023 **Published Online:** 11 January 2024

12 January 2024 09:05:23

1. Introduction

When talking in a noisy environment, speakers adjust their output in several ways, resulting in so-called Lombard speech. In addition to speaking more loudly, they increase their fundamental frequency (f_0; Bond *et al.*, 1989; Ibrahim *et al.*, 2022; Junqua, 1993; Lunichkin *et al.*, 2023; Summers *et al.*, 1988) and increase energy in higher harmonics of the produced speech signal (Junqua, 1993; Summers *et al.*, 1988), suggesting possible adaptations of voice source characteristics. Lombard speech also shows a higher ratio between vowels and consonants in terms of duration and intensity; vowels tend to be even longer and even louder than consonants (Castellanos *et al.*, 1996; Junqua, 1993). These adaptations of voice source and temporal features are accompanied by spatial adjustments of supraglottal articulation, such as larger lip aperture and jaw lowering (Garnier *et al.*, 2006; Garnier *et al.*, 2018; Scobbie and Ma, 2019a; Šimko *et al.*, 2016; Trujillo *et al.*, 2021), and lingual hyper-articulation (Šimko *et al.*, 2016; Trujillo *et al.*, 2021), resulting in an increase in *F*1 (Bond *et al.*, 1989; Garnier *et al.*, 2006; Garnier *et al.*, 2018; Ibrahim *et al.*, 2022; Junqua, 1993; Lunichkin *et al.*, 2023; Scobbie and Ma, 2019b; Summers *et al.*, 1988). These acoustic and articulatory modifications are hypothesized to arise to improve the auditory feedback of the speaker in noisy environments (Luo *et al.*, 2018) and/or increase speech intelligibility for the listener (Garnier *et al.*, 2010; Garnier *et al.*, 2018; Junqua *et al.*, 1999), including articulatory enhancements of visual cues (Alexanderson and Beskow, 2014; Fitzpatrick *et al.*, 2015; Garnier *et al.*, 2018; Trujillo *et al.*, 2021).

In this paper, we model articulatory effects of Lombard speech, namely, spatial adjustments of lip and jaw articulation and *F*1 increase, as emergent consequences of acoustic adaptations to maintain a reasonable level of intelligibility of produced speech. This model of speech production is based on the optimal control theoretic (OCT; Todorov, 2006) assumption that speech articulation satisfies conflicting requirements of production efficiency and perceptual efficacy (Elie *et al.*, 2023a,b; Nelson, 1983; Parrell and Lammert, 2019; Patri *et al.*, 2015; Simko and Cummins, 2010; Windmann *et al.*, 2015). Articulatory movements are, thus, predicted to minimize a multi-objective cost function, modeled as a weighted sum of cost components associated with production effort and decreases in intelligibility for a listener. This idea follows Lindblom's H&H theory (Lindblom, 1990) that variations associated with hypo- and hyper-articulation emerge from the balance between the effort cost and intelligibility requirement (Lindblom, 1990). The model presented here builds on previous optimal control theory studies where speech intelligibility was conceptualized as a function of the acoustic correlates of phonemes (Elie *et al.*, 2023a,b), surface duration of speech constituents (Windmann *et al.*, 2015), or the distance from articulatory targets (Simko and Cummins, 2010, 2011; Šimko *et al.*, 2014).

The present account offers an extension of the OCT-based model of speech production proposed in Elie *et al.* (2023a) by adding a term which compensates for the loss of intelligibility due to decreasing *acoustic* signal-to-noise ratio in the intelligibility model. Using an articulatory model (Maeda, 1990) that maps articulatory configurations to acoustic

[a]Author to whom correspondence should be addressed.

 ©Author(s) 2024.

consequences, we present simulations predicting articulatory configurations of vowels for various levels of background noise, based purely on the requirements of optimality.

Section 2 presents the OCT-based model of speech production. Section 3 presents the acoustic model used to compute the acoustic intensity, and Sec. 4 presents our extension of the intelligibility model that accounts for loss of intelligibility in noisy environments. Finally, Sec. 5 presents the simulations of vocalic configurations produced in various levels of background noise. Simulations are intended to compare the results of our model with observed articulatory and acoustic effects of Lombard speech. Given the simplistic nature of our model, this comparison is purely qualitative.

2. Optimal control of speech

The multi-objective cost function used in this paper models two costs: a least effort cost and a cost of not being intelligible, hence, the following cost function:

$$\mathcal{C}(\mathbf{x}) = \alpha_{\mathcal{E}}\mathcal{E}(\mathbf{x}) + \alpha_{\mathcal{I}}(1 - \mathcal{I}(\mathbf{x})), \tag{1}$$

where $\mathcal{E}$ and $\mathcal{I}$ are the effort and intelligibility functions, respectively, and $\alpha_{\mathcal{E}}$ and $\alpha_{\mathcal{I}}$ are the weights assigned to the effort cost and intelligibility requirement, respectively. The vector $\mathbf{x}$ contains the parameters to optimize, namely, a vector of static articulatory parameters.

Similar to our previous paper (Elie *et al.*, 2023b), here, we consider only static configurations using the Maeda articulatory model (Maeda, 1990). Consequently, the vector $\mathbf{x}$ from Eq. (1) contains the values of the seven parameters of the Maeda model, where each value is contained between -3 and $+3$.

Additionally, we use the same model of articulatory effort as was used in Elie *et al.* (2023b), namely, a function of the normalized Euclidean between the target distance of the target articulatory position and their position at rest:

$$E(\mathbf{x}) = \frac{1}{E_{\max}}||\mathbf{x}||^2 = \frac{1}{63}||\mathbf{x}||^2, \tag{2}$$

where $E_{\max} = 63$ is used to normalize the effort cost to a value between zero and one, leading to the same value range as intelligibility.

3. The acoustic model

The acoustic model presented in this section is used to compute formant frequencies and also compute sound pressure level, defined as in Eq. (3). Formant frequencies are computed using the Maeda model (Maeda, 1990), which can generate midsagittal shapes of the vocal tract from a vector of seven independent articulatory parameters.

Intensity was computed from acoustic simulations of speech given a static configuration of the vocal tract using extended single matrix formulation (ESMF) synthesis (Elie and Laprie, 2016). ESMF is a physical model of the vocal tract which simulates acoustic propagation in time-varying vocal tract area functions. It is coupled with a two-mass self-oscillating model of the vocal folds. We used static area functions to simulate 250 ms of speech. Subglottal pressure and fundamental frequency of the glottal source were fixed at 800 Pa and 100 Hz, respectively. Sound pressure level was computed as

$$L_{\mathrm{dB}} = 20\ \log_{10}(p_{\mathrm{rms}}/p_{\mathrm{ref}}), \tag{3}$$

where p_{rms} is the root mean square value of the output pressure signal and $p_{\mathrm{ref}} = 2 \times 10^{-5}$ Pa is the reference sound pressure. Figure 1 shows examples of simulated speech signals for two French vowels /ä/ and /i/. This example shows that given a fixed subglottal pressure and a fixed fundamental frequency, the output acoustic level may change depending on the shape

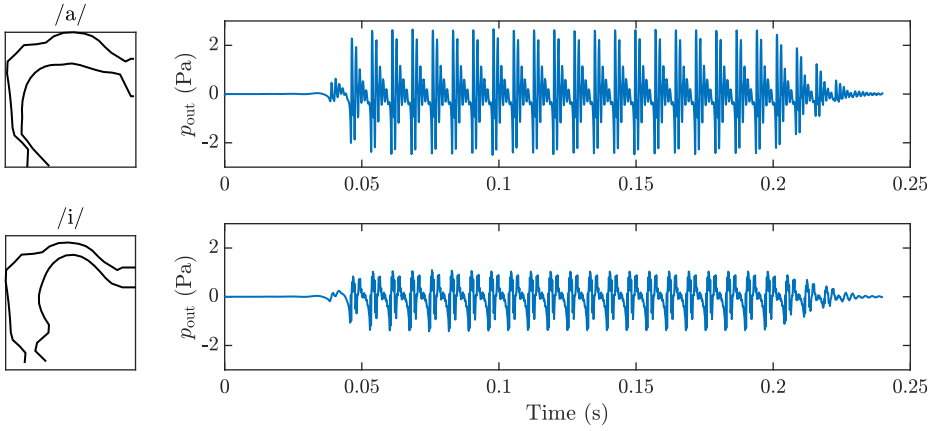

Fig. 1. Example of simulated speech sounds for two static vowel configurations, corresponding to /ä/ (top) and /i/ (bottom). (Left side) Midsagittal shape of the vowels returned by the Maeda model, and (right side) simulated sound pressure are shown.

12 January 2024 09:05:23

of the vocal tract. The sound pressure level, computed as in Eq. (3), is 93.8 dB for /ä/ and 88.5 dB for /i/. Note that p_{rms} is computed over an interval included in the steady state of the acoustic pressure (here, between $t = 0.075$ s and $t = 0.175$ s).

4. The intelligibility model

In this paper, the intelligibility model is derived from two different aspects: vowel formant frequencies and sound pressure level. For vowel formant frequencies, the intelligibility model is similar to the model presented in our recent paper (Elie *et al.*, 2023b), i.e., a simple model of the probability of vowel recognition based solely on its formant frequencies. In this paper, we modulate the intelligibility function as a function of signal-to-noise ratio, such as

$$\mathcal{I} = P(v|\mathbf{f}) \times g(\Delta L), \tag{4}$$

where $P(v|\mathbf{f})$ is the conditional probability of recognition of the vowel, v, given the formant vector, $\mathbf{f}$, and $g(\Delta L)$ is a function modeling the loss of intelligibility as a function of the signal-to-noise ratio, ΔL. The signal-to-noise ratio, ΔL, is defined as the difference in dB between the sound pressure level of the produced vowel and the sound pressure level of the background noise.

4.1 Intelligibility based on formant frequencies

The conditional probability, $P(v|\mathbf{f})$, of vowel, v, recognition given a formant vector, $\mathbf{f}$, is computed using a specifically designed formant-to-probability (FtP) model. The FtP model used in this paper was built following an approach similar to the one presented in Elie *et al.* (2023a,b). The main difference is the use of a quadratic discriminant analysis (QDA) model instead of a Gaussian mixture model (GMM). Using a QDA enabled the nonvocalic samples (#) to be grouped in a single class, as opposed to the GMM which requires modeling them into a large number of Gaussians to cover their distribution in the acoustic space.

Additionally, instead of using formants extracted from an American English database, we used formants extracted from a French database, the IFCASL (Trouvain *et al.*, 2016) oral corpus, which contains read speech of 54 French native speakers. We chose to use French vowels to make a qualitative comparison possible with previous Lombard speech studies based on French data (Garnier *et al.*, 2006; Garnier *et al.*, 2018). The QDA was trained to recognize 11 monophthong vowels of French.

4.2 Intelligibility based on acoustic intensity

For this paper, we consider a global function of intelligibility as a function of the signal-to-noise ratio, ΔL, which is applied independently of vowel category. For that purpose, the loss of intelligibility function, $g(\Delta L)$, has been approximated by evaluating the intelligibility of utterances as a function of the signal-to-noise ratio using automatic speech recognition (ASR). It consisted of computing the word error rate (WER) for real speech, returned by the ASR system Whisper (Radford *et al.*, 2022) to which a controlled acoustic level of white noise had been artificially added. We used the same corpus for the real speech signals as for the training data of the FtP model described in Sec. 4.1. The level of noise was controlled such that the actual signal-to-noise ratio, ΔL, matched a target, which was varied between −10 and 50 dB with an increment step of 1 dB. We used the pretrained large-v2 ASR model from Whisper (Radford *et al.*, 2022) with language set to French.

Figure 2 represents the WER returned by the ASR system Whisper as a function of ΔL. It shows a sigmoid function where, as expected, the WER is high when the signal-to-noise ratio, ΔL, is low and the WER is low for high ΔL. The next step consisted of fitting a logistic function to these observations in the form

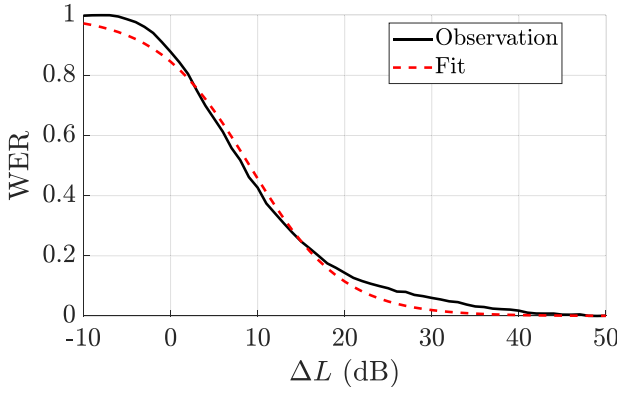

Fig. 2. WER as a function of signal-to-noise ratio, ΔL. The solid line represents the observation and the dashed line represents the fitted logistic function as defined in the text.

12 January 2024 09:05:23

$$\sigma(\Delta L) = 1 - \frac{1}{1 + e^{-\beta(\Delta L - \gamma)}}, \tag{5}$$

where β and γ are the two parameters to estimate. The fit has been performed by minimizing the root mean square error between the generated σ function and the observation. Figure 2 represents the fitted σ function as a dashed line with $\beta = 0.1873$ and $\gamma = 9.0783$. In our model, the g function should increase with ΔL as a speech sound is more likely to be recognized in a high signal-to-noise ratio than in a low one. Consequently, we chose to define $g(\Delta L)$ to apply in the intelligibility function of Eq. (4) as

$$g(\Delta L) = 1 - \sigma(\Delta L) = \frac{1}{1 + e^{-\beta(\Delta L - \gamma)}}. \tag{6}$$

5. Simulations

In this section, we present the results of simulations consisting of optimizing the three cardinal vowels of the French vocalic triangle, namely, /ä, i, u/ for various background noise levels. In all of these simulations, the weight assigned to the intelligibility cost, $\alpha_{\mathcal{I}}$, is set to one.

The optimization method was as follows. We started from a randomly generated initial solution, where each parameter of the Maeda vector followed a uniform distribution between -3 and $+3$. Then we used the Nelder-Mead method (Nelder and Mead, 1965) to find a local minimum. Once the Nelder-Mead method found a local minimum, we modified the solution randomly and reran the Nelder-Mead optimization. We repeated this step until a convergence criterion was met: if the new local solutions did not improve the best solution for three successive attempts, we returned the global best solution. For each simulation, we ran this process 150 times and kept the solution for which the cost function was the lowest.

5.1 Results with constant weight assigned to least effort

In a first set of simulations, the weight assigned to the least effort requirement, $\alpha_{\mathcal{E}}$, is kept constant, either zero or one, for various background noise levels. The background noise level is increased from 0 to 90 dB with an increment step of 10 dB.

Figure 3 shows the results with a constant weight assigned to least effort for two different values of $\alpha_{\mathcal{E}}$, namely, zero and one. When $\alpha_{\mathcal{E}} = 0$, there is no constraint on articulatory effort: the optimization consists, then, of finding the solution which maximizes intelligibility as defined in Eq. (4), regardless of the articulatory effort required to produce this solution. When $\alpha_{\mathcal{E}} = 1$, articulatory effort is penalized: optimization returns a solution which is a balance between high intelligibility and low articulatory effort. When no constraint is applied to the least effort requirement, namely, when $\alpha_{\mathcal{E}} = 0$, we observe (top left panel of Fig. 3) a shrinking of the vowel space: $F1$ increases for /i/ and /u/ with the background noise level at it goes from around 200 Hz for $L_{\text{noise}} = 0$ dB to around 400 Hz for $L_{\text{noise}} = 90$ dB while it decreases

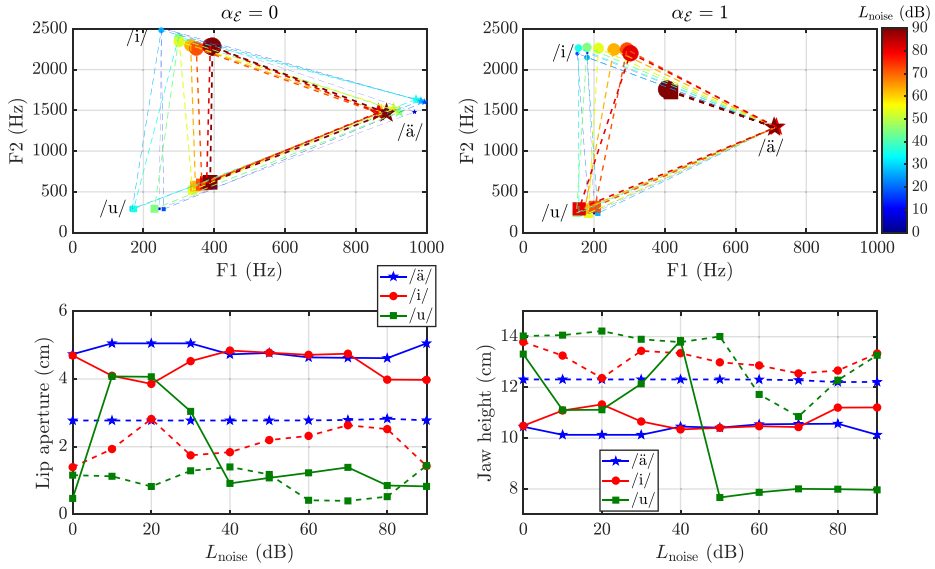

Fig. 3. Results of simulations with a constant least effort constraint. (Top) The positions of the simulated vowels /ä, i, u/ are shown in the $F1$-$F2$ vowel space for different levels of background noise (from 0 to 90 dB) for no least effort requirement ($\alpha_{\mathcal{E}} = 0$, top left) and with $\alpha_{\mathcal{E}} = 1$ (top right). (Bottom) The lip aperture (left) and the jaw height (right) are shown for the simulated vowels as a function of background noise. The lip apertures and jaw heights obtained for $\alpha_{\mathcal{E}} = 0$ are denoted by solid lines while dashed lines denote lip apertures and jaw heights obtained for $\alpha_{\mathcal{E}} = 1$.

12 January 2024 09:05:23

for /ä/ (going from 1000 Hz for $L_{noise} = 0$ dB to less than 900 Hz for $L_{noise} = 90$ dB). For the three vowels, there is no significant change in lip aperture with various background noise levels. In this case, our model did not succeed in predicting the articulatory and acoustic adjustments associated with Lombard effect, such as more lip aperture (Garnier *et al.*, 2006; Garnier *et al.*, 2018; Scobbie and Ma, 2019a; Šimko *et al.*, 2016; Trujillo *et al.*, 2021) and raise of *F*1 in low vowels (Bond *et al.*, 1989; Garnier *et al.*, 2006; Garnier *et al.*, 2018; Ibrahim *et al.*, 2022; Junqua, 1993; Lunichkin *et al.*, 2023; Scobbie and Ma, 2019b; Summers *et al.*, 1988). When the simulations do not consider any constraint on articulatory effort in low-level background noise (the dark purple line in the top left panel in Fig. 3), simulations return a purely intelligibility-optimized solution. However, in real speech situations, this purely intelligibility-based optimization of speech is not necessary in a quiet environment, and speakers are more likely to additionally use the least effort constraint to determine their degree of hypo-/hyper-articulation in speech (Lindblom, 1990). It is possible that these purely intelligibility-optimized solutions also correspond to or are close to the sound-pressure-level-optimized solutions. This is supported by the fact that at equivalent signal-to-noise ratio, Lombard speech has been reported to be more intelligible than non-Lombard speech (Dreher and O'Neill, 1957; Summers *et al.*, 1988).

Figure 3 also shows the results when articulatory effort is penalized with $\alpha_{\mathcal{E}} = 1$. The top right plot of Fig. 3 shows a significant acoustic effect on /i/, which tends to centralize: *F*1 and *F*2 move toward the center of the vocalic triangle. Conversely, there is no significant effect on *F*1 and *F*2 for /ä/ and /u/. Note that /u/ is completely centralized for very high noise level (90 dB) and merges with /i/. For /i/, we observe an increase in lip aperture with increased background noise level, starting from around 1.5 cm for $L_{noise} = 0$ dB, up to around 2.5 cm for $L_{noise} = 80$. Then, lip aperture drops to around 1.5 cm for $L_{noise} = 90$ dB. One possible explanation for the drop in lip aperture at 90 dB is that to compensate for the loss in intelligibility for this background noise level would require too much articulatory effort; the optimal result in this case is a vocal tract configuration close to the neutral configuration, hence, centralization. This would also explain the centralization of /u/. Similar to the case without an effort cost, lip aperture for /ä/ stays constant. However, note that it is now slightly lower than 3 cm, which is less than that in the no least effort requirement case, where a lip aperture of /ä/ of around 5 cm was predicted. A likely explanation is that producing a large lip aperture of i.e., 5 cm, requires too much effort. As a consequence, when effort is penalized, the optimal /ä/ is located closer to the center of the vocalic space as shown in the top panel of Fig. 3.

These preliminary experiments showed that our model fails to predict the articulatory and acoustic effects of Lombard speech when the speaker keeps the effort cost weight constant for different signal-to-noise ratios. This result suggests that the speaker needs to adapt their degree of hypo- and hyper-articulation of speech to activate Lombard speech. This idea is explored in Sec. 5.2, where simulations are presented with varying weights assigned to the effort cost for various background noise levels.

5.2 Results with decreasing weight assigned to the effort cost for increasing background noise

In a second set of simulations, the weight assigned to the effort cost, $\alpha_{\mathcal{E}}$, varies as a function of background noise level. We chose to apply a simple linear function of the background noise level L_{noise} as follows:

$$\alpha_{\mathcal{E}} = -\frac{1}{50}(L_{noise} - 40) + 1. \tag{7}$$

This is designed with $\alpha_{\mathcal{E}} = 1$ for $L_{noise} = 40$ dB and $\alpha_{\mathcal{E}} = 0$ for $L_{noise} = 90$ dB. Similar to the first set of simulations presented in Sec. 5.1, background noise level is increased from 0 to 90 dB with an increment step of 10 dB.

Figure 4 shows results when the weight assigned to the effort cost follows Eq. (7). In this case, our model predicts the vocalic shift previously observed in real Lombard speech (Garnier *et al.*, 2006; Garnier *et al.*, 2018): the first formant, *F*1, is raised in frequency as the background level increases. In addition, our model also predicts a slight increase in *F*2 with background noise level. For /u/ and /ä/, the effect of Lombard speech is visible when the background noise is above a certain level. Indeed, the vocalic shift appears when the background noise is above 70 dB for /ä/ and 80 dB for /u/. Interestingly, the effect is much more progressive for /i/; in this case, *F*1 and *F*2 increase gradually with background noise level.

The articulatory effect of Lombard speech, as shown in the bottom panel of Fig. 4, reflects the acoustic effect for /ä/ and /i/. Lip aperture for /ä/ increases when background noise is above 80 dB while lip aperture for /i/ increases continuously with background noise level. Jaw height shows the opposite pattern, which suggests that the increase in lip aperture is mainly caused by the lowering of the jaw. For /u/, although we observe a vocalic shift similar to the other vowels, we do not observe a modification of lip aperture as it stays rather constant around 1 cm. However, we observe a significant lowering of the jaw; jaw height goes from 14 cm for $L_{noise} = 0$ dB to less than 8 cm for $L_{noise} = 90$ dB. One possible explanation for this is that French /u/ requires strong lip protrusion, which would prevent large lip aperture. However, a larger lip aperture for Lombard /u/ has been experimentally observed in French (Garnier *et al.*, 2018).

6. Conclusion and future work

The presented work serves as a proof of concept. The simulations successfully showed that an optimal control theory model, which includes a cost for effort and the requirement of being intelligible, can predict several key articulatory characteristics of Lombard speech and their acoustic consequences. Increasing background noise level has the effect of raising

12 January 2024 09:05:23

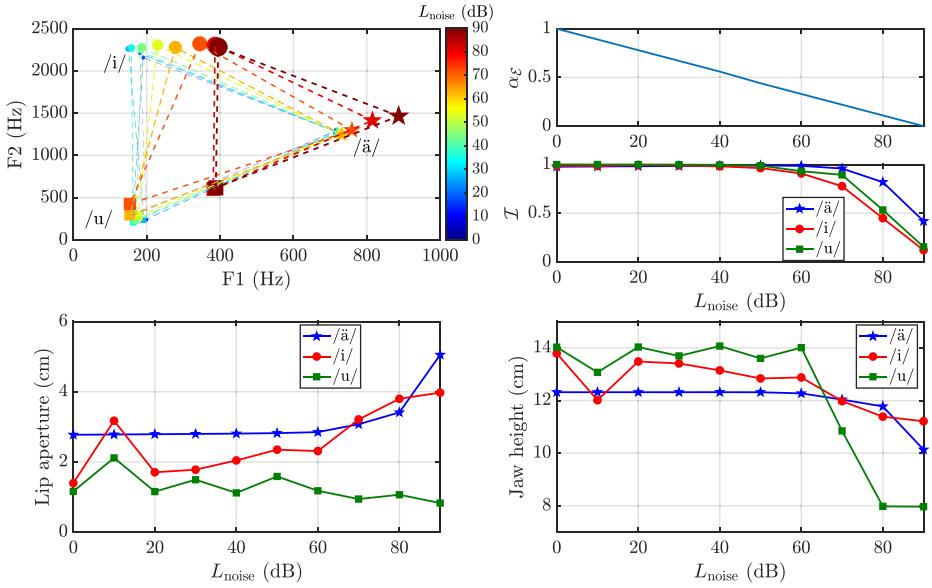

Fig. 4. Results of our simulations with a decreasing weight assigned to the effort cost for increasing background noise. The top left plot shows the position of the simulated vowels /ä, i, u/ in the *F*1-*F*2 vowel space for different levels of background noise (from 0 to 90 dB). The top right panel shows the weight assigned to the least effort requirement, $\alpha_\mathcal{E}$, as a function of the background noise level, L_{noise}, following Eq. (7). The middle right plot shows the intelligibility function, $\mathcal{I}$, corresponding to the optimal solution for the three vowels as a function of the noise level, L_{noise}. (Bottom) The lip aperture (left) and jaw height (right) for the simulated vowels as a function of background noise.

the first (and second) formant frequency and lowering the jaw for the simulated vowels /ä/, /i/, and /u/, as well as increasing lip aperture in the optimal solutions for /ä/ and /i/ but not for /u/. The vowel /u/ also showed significant jaw lowering in the presence of a high level of background noise. These results are consistent with observed effects in the literature (Bond *et al.*, 1989; Garnier *et al.*, 2006; Garnier *et al.*, 2018; Ibrahim *et al.*, 2022; Lunichkin *et al.*, 2023; Scobbie and Ma, 2019a,b; Šimko *et al.*, 2016).

For the very high volume of background noise, the articulation of (acoustically sufficiently similar) /u/ seems to have switched to a very low (open) jaw position while maintaining small lip aperture. Although optimal in terms of the constraints imposed by our OCT account, this articulatory configuration might be difficult for a human vocal tract to produce. To mitigate this discrepancy, more realistic "embodiment" constraints might need to be imposed on the static articulatory model in the future.

To obtain realistic articulatory effects, we needed to release the constraints applied to articulatory effort in the presence of noise. Within the optimal control theory framework, this suggests that when producing Lombard speech, speakers adapt their speaking style, namely, they allow for a more effortful articulation compared to speaking in a more quiet environment. This is consistent with previous studies which showed that for an equivalent signal-to-noise ratio, Lombard speech is more intelligible than speech produced in a quiet environment (Dreher and O'Neill, 1957; Summers *et al.*, 1988): intelligibility is given more importance than the articulatory effort in Lombard speech.

Importantly, the modeling predictions presented in this paper can be treated as emergent from balancing the requirements of production efficiency and perceptual efficacy of speech produced in the loud environment. The presented model does not contain any requirements explicitly driving the resulting articulation toward the observed patterns; these are instead consequences of trade-offs between independently motivated objectives encompassing the speaker and the listener. Also, as we use purely listener-oriented intelligibility criteria as a relevant constraint, these modeling results lend support to the listener-oriented hypothesis of Lombard speech (Garnier *et al.*, 2010; Garnier *et al.*, 2018; Junqua *et al.*, 1999).

We consider this paper as a first step in optimization-based modeling of speech production and variation due to background intensity level adjustments; therefore, the presented results are primarily qualitative and may not reproduce quantitative details of adjustments. Nevertheless, despite many simplifications, the model reproduces several key observed Lombard speech phenomena.

In particular, some of the supraglottal adjustments in Lombard speech can be expected to interact with adjustments in voice source features, including subglottal pressure, as well as acoustic correlates of changes in glottal source characteristics, i.e., changes in f_0, and changes in spectral tilt, which we have not included in our current model. We aim to implement these adjustments—and their consequences in terms of intelligibility and articulatory effort—in future versions of our model. One of the simplifications that we make is the assumption that the intelligibility function presented in Fig. 2 is independent of vowel quality and voice source features (such as sub-glottal pressure, f_0, and spectral tilt).

12 January 2024 09:05:23

Also, Eq. (4) assumes that the influences of formant frequencies and signal-to-noise related phenomena on intelligibility do not interact in a complex way. The potential interactions between the different sources of intelligibility loss may quantitatively impact the adjustments in different ways for different vowels shown in Figs. 3 and 4. However, we expect the qualitative pattern to hold for all vowels, i.e., an increase in lip aperture, jaw lowering, and consequent *F*1 increase for increased levels of background noise.

Another area of future development will focus on adapting the present account to the dynamics of speech articulation using dynamical models, such as those presented recently (Elie *et al.*, 2023a). These, we believe, will enable us to model effects such as faster jaw movements (Šimko *et al.*, 2016) in Lombard speech or larger modifications of duration and intensity of vowels than of consonants (Castellanos *et al.*, 1996; Junqua, 1993).

Acknowledgment

We gratefully acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (PlanArt: Planning the Articulation of Spoken Utterances; ERC Advanced Grant awarded to A.T., Grant No. 101019847).

Author declarations

Conflict of Interest

All authors declare that they have no conflicts of interest to disclose.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Alexanderson, S., and Beskow, J. (**2014**). "Animated Lombard speech: Motion capture, facial animation and visual intelligibility of speech produced in adverse conditions," Comput. Speech Lang. **28**(2), 607–618.

Bond, Z., Moore, T. J., and Gable, B. (**1989**). "Acoustic–phonetic characteristics of speech produced in noise and while wearing an oxygen mask," J. Acoust. Soc. Am. **85**(2), 907–912.

Castellanos, A., Benedí, J.-M., and Casacuberta, F. (**1996**). "An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect," Speech Commun. **20**(1-2), 23–35.

Dreher, J. J., and O'Neill, J. (**1957**). "Effects of ambient noise on speaker intelligibility for words and phrases," J. Acoust. Soc. Am. **29**(12), 1320–1323.

Elie, B., and Laprie, Y. (**2016**). "Extension of the single-matrix formulation of the vocal tract: Consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink," Speech Commun. **82**, 85–96.

Elie, B., Šimko, J., and Turk, A. (**2023a**). "Optimal control of speech with context-dependent articulatory targets," in *Interspeech 2023*, Dublin, Ireland (ISCA, Dublin, Ireland).

Elie, B., Šimko, J., and Turk, A. (**2023b**). "Optimal control theory of speech production using probabilistic articulatory-acoustic models," in *20th International Conference of Phonetic Sciences (ICPhS)*, Prague, Czech Republic (Guarant International, Prague, Czech Republic).

Fitzpatrick, M., Kim, J., and Davis, C. (**2015**). "The effect of seeing the interlocutor on auditory and visual speech production in noise," Speech Commun. **74**, 37–51.

Garnier, M., Bailly, L., Dohen, M., Welby, P., and Loevenbruck, H. (**2006**). "An acoustic and articulatory study of Lombard speech: Global effects on the utterance," in *Ninth International Conference on Spoken Language Processing*, Pittsburgh, PA (ISCA, Pittsburgh, PA.

Garnier, M., Bernardoni, N. H., and Dubois, D. S. (**2010**). "Influence of sound immersion and communicative interaction on the Lombard effect," J. Speech. Lang. Hear. Res. **53**(3), 588–608.

Garnier, M., Ménard, L., and Alexandre, B. (**2018**). "Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues?," J. Acoust. Soc. Am. **144**(2), 1059–1074.

Ibrahim, O., Yuen, I., van Os, M., Andreeva, B., and Möbius, B. (**2022**). "The combined effects of contextual predictability and noise on the acoustic realisation of German syllables," J. Acoust. Soc. Am. **152**(2), 911–920.

Junqua, J.-C. (**1993**). "The Lombard reflex and its role on human listeners and automatic speech recognizers," J. Acoust. Soc. Am. **93**(1), 510–524.

Junqua, J.-C., Fincke, S., and Field, K. (**1999**). "The Lombard effect: A reflex to better communicate with others in noise," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP99)*, Phoenix, AZ (IEEE, New York), Vol. 4, pp. 2083–2086.

Lindblom, B. (**1990**). "Explaining phonetic variation: A sketch of the H&H theory," in *Speech Production and Speech Modelling* (Springer, Dordrecht), pp. 403–439.

Lunichkin, A. M., Andreeva, I. G., Zaitseva, L. G., Gvozdeva, A. P., and Ogorodnikova, E. A. (**2023**). "Changes in the spectral characteristics of vowels in Russian speech on a noise background," Acoust. Phys. **69**, 357–366.

Luo, J., Hage, S. R., and Moss, C. F. (**2018**). "The Lombard effect: From acoustics to neural mechanisms," Trends Neurosci. **41**(12), 938–949.

Maeda, S. (**1990**). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modelling* (Springer, Dordrecht), pp. 131–149.

Nelder, J. A., and Mead, R. (**1965**). "A simplex method for function minimization," Comput. J. **7**(4), 308–313.

Nelson, W. L. (**1983**). "Physical principles for economies of skilled movements," Biol. Cybern. **46**(2), 135–147.

Parrell, B., and Lammert, A. C. (**2019**). "Bridging dynamical systems and optimal trajectory approaches to speech motor control with dynamic movement primitives," Front. Psychol. **10**, 2251.

12 January 2024 09:05:23

Patri, J.-F., Diard, J., and Perrier, P. (**2015**). "Optimal speech motor control and token-to-token variability: A Bayesian modeling approach," Biol. Cybern. **109**, 611–626.
Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (**2022**). "Robust speech recognition via large-scale weak supervision."
Scobbie, J. M., and Ma, J. (**2019a**). "Say again? Individual articulatory strategies for producing a clearly-spoken minimal pair wordlist," in *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*, Melbourne, Australia (Australasian Speech Science and Technology Association, Inc., Canberra, Australia).
Scobbie, J. M., and Ma, J. (**2019b**). "Say again? Individual acoustic strategies for producing a clearly-spoken minimal pair wordlist," in *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*, Melbourne, Australia (Australasian Speech Science and Technology Association, Inc., Canberra, Australia).
Šimko, J., Beňuš, Š., and Vainio, M. (**2016**). "Hyperarticulation in Lombard speech: Global coordination of the jaw, lips and the tongue," J. Acoust. Soc. Am. **139**(1), 151–162.
Simko, J., and Cummins, F. (**2010**). "Embodied task dynamics," Psychol. Rev. **117**(4), 1229–1246.
Simko, J., and Cummins, F. (**2011**). "Sequencing and optimization within an embodied task dynamic model," Cognit. Sci. **35**(3), 527–562.
Šimko, J., O'Dell, M., and Vainio, M. (**2014**). "Emergent consonantal quantity contrast and context-dependence of gestural phasing," J. Phonet. **44**, 130–151.
Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (**1988**). "Effects of noise on speech production: Acoustic and perceptual analyses," J. Acoust. Soc. Am. **84**(3), 917–928.
Todorov, E. (**2006**). "Optimal control theory," in *Bayesian Brain: Probabilistic Approaches to Neural Coding*, edited by K. Doya (MIT Press, Cambridge, MA), pp. 268–298.
Trouvain, J., Bonneau, A., Colotte, V., Fauth, C., Fohr, D., Jouvet, D., Jügler, J., Laprie, Y., Mella, O., Möbius, B., and Zimmerer, F. (**2016**). "The IFCASL Corpus of French and German non-native and native read speech," in *Proceedings LREC'2016, 10th Edition of the Language Resources and Evaluation Conference*, Portorož, Slovenia (European Language Resources Association, Portoroz, Slovenia).
Trujillo, J., Özyürek, A., Holler, J., and Drijvers, L. (**2021**). "Speakers exhibit a multimodal Lombard effect in noise," Sci. Rep. **11**, 16721.
Windmann, A., Šimko, J., and Wagner, P. (**2015**). "Optimization-based modeling of speech timing," Speech Commun. **74**, 76–92.

12 January 2024 09:05:23